Annotation of Multiword Expressions in the Farsi Section of the Universal Dependencies Project

Behrang QasemiZadeh
Heinrich-Heine-Universität Düsseldorf

1 Summary

Inspired by the work presented by De Smedt et al. (2015), the annotation of the multiword expressions (MWEs) in the Farsi section of the universal dependencies (UD) project[1] is reviewed. To do so, the annotation of similar syntactic structures in the Farsi and English sections of the UD project are compared using the *INESS-Search system*[2] (similar as described in De Smedt et al. (2015)). A number of observations from this initial study are reported.

2 Farsi, the Modern Persian Language

Farsi, also known as Persian, is the official language of Iran and it is spoken by about 100 million people. Farsi belongs to the family of Indo-European languages and has straightforward morphology (QasemiZadeh and Rahimi, 2006) and syntax (Seraji, 2015): Derivation and inflection are carried out using affixation; case markers are rarely used and the word order is not restricted (although the SOV pattern is dominant). While these proprieties make Farsi easy to learn and marvellous to speak, they can also make it hard for modelling by machines.

Further complications in the formal modelling and analysis of Farsi is caused by the fact that it is written using the Arabic transliteration system. Due to the use of Arabic transliteration,[3] identifying the boundaries of words can be problematic; put simply, white spaces may not represent the boundaries of words.

Another peculiar characteristic of Farsi, which can be potentially interesting for the study of MWEs, is the way that actions and events are described: The number of main/simple verbs (i.e., single token) in Farsi is extremely limited—in major Farsi dictionaries, the number of such entries is less than 800. Instead, actions and events are described using compound verbs. Most works (including the Farsi section of UD project) traditionally limit the syntactic structure of verb compounds to the combination of a 'light verb' and a noun, adjective, prepositional phrase, adverb, or past participle. However, as discussed in details by Dabir-Moghaddam (1997), verbs other than light are also lexicalised in a similar fashion as light verbs to construct compound verbs. The latter, however, is often neglected by syntactician due various reasons (see Dabir-Moghaddam, 1997).

3 Comparing Annotations of MWEs for Farsi and English

UD project still lacking consistent annotations for relations that deal with MWEs (i.e., `compound` and its sub-relations, `mwe`, and `name`).[4] When browsing the Farsi corpus, at first sight, it seems that the above-mentioned relations are introduced irrespective of semantics and with the sole focus on filling the void caused by the lack of any syntactic relationship between the elements of a MWEs. In a number of circumstances, this policy has resulted in conflicting annotations between comparable English and Farsi structures.

[1] `https://universaldependencies.github.io/docs/`, ver 1.1; McDonald et al. (2013).

[2] `http://clarino.uib.no/iness/page`

[3] Note that Arabic transliteration is designed for a Semitic language with a template-based morphology.

[4] Note that documentations for Farsi are still missing from the UD ver 1.3. Also note that the annotations in UD project are evolving and issues that are named here may be corrected in the future releases.

- compound:
  - For Farsi, the language specific relation compound:lvc for annotating light verb construct is introduced. For the reasons mentioned by linguists such as Dabir-Moghaddam (see 1997), I believe that compound:lvc is not sufficient for the annotation of all compound verbs in Farsi.[5]
  - In the English corpus, compound relation is employed to mark syntactic relations in transliterated numbers and connect elements such as 'thousand' or 'million' to numbers. In Farsi, nummod relation replaces compound to connect elements such as 'thousand' or 'million' to numbers that are rightly connected to each other using cc relation.[6] Interestingly, in Farsi corpus, compound is employed to mark syntactic relations that are often marked as nummod in the English corpus—for example, the relation between 'two' and 'pictures' in 'two pictures'.
  - In the English corpus, the elements of many multi-word specialised vocabularies (such as the name of organisations, financial terms, etc.) are connected by the compound relation. For example, in the English corpus, for the term 'search engine', 'search' is the dependant of 'engine' using compound. In the Farsi corpus, comparable structures are marked using nmod and nmod:poss relations. While it can be argued that these annotations in the Farsi corpus are syntactically correct, they are not consistent with the English corpus's annotation.
- name
  - name is used similarly for both Farsi and English. However, English honorifics are connected to their regents using compound; in Farsi, honorifics are connected to their regents using the name relation.
- mwe
  - Similar to the English corpus, in the Farsi corpus, mwe is mostly employed to mark prepositional compounds—particularly those that end to the conjunction */ke/* (conjunction */ke/* plays a comparable role as 'that' in English). The use of mwe, however, is somehow confusing. Annotating every sequence of prepositions that ends with conjunction */ke/* (which can show a high degree of productivity) looks an overuse of the relation mwe and inconsistent with the English annotations (in which 'that' is related using the relation case). Although these annotations in the Farsi corpus are not wrong (at the end, these are all definitions), they could be more consistent with the English ones (e.g., by following the given definition for the relation mark:marker for English).
  - Simultaneously, mwe seems to be an underused relation in the Farsi corpus. For example, mwe many prepositional expressions, which are annotated in the English corpus using mwe, are annotated in the Farsi corpus using the case relation.

References

Dabir-Moghaddam, M. (1997). Compound verbs in Persian. *Studies in the Linguistic Sciences*, 27(2).

De Smedt, K., Rosen, V., and Meurer, P. (2015). Studying consistency in UD treebanks with INESS-Search. In *Proceedings of TLT14*, pages 258–267. Polish Academy of Sciences.

McDonald, R., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., Hall, K., Petrov, S., Zhang, H., Täckström, O., Bedini, C., Bertomeu Castelló, N., and Lee, J. (2013). Universal dependency annotation for multilingual parsing. In *Proceedings of ACL (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria. ACL.

QasemiZadeh, B. and Rahimi, S. (2006). Persian in MULTEXT-East framework. In *Advances in Natural Language Processing*, volume 4139 of *LNCS*, pages 541–551. Springer.

Seraji, M. (2015). *Morphosyntactic Corpora and Tools for Persian*. Acta Universitatis Upsaliensis.

[5]Note that the current annotation in the Farsi corpus has a number of other issues with verbal structures, for example, in many cases, while the active voice of a verb is annotated using compound:lvc, the passive voice of the same verb is not.

[6]In Farsi, numbers are connected using the conjunction /va/.